F$_1$ macro scores (100 test points, GPT-3 fine tuned on SMILES)